CLAIMS

What is claimed is:

1. A method for speech-based information retrieval in Mandarin Chinese, comprising:

entering voice or text queries describing information to be requested;

determining the indexing terms; and

using the indexing terms to retrieve the information records requested in a format of voice or text type,

wherein the indexing terms are overlapping syllable segments with a specific length, and the specific length can be assigned arbitrarily and is at least one.

2. The method for speech-based information retrieval in Mandarin Chinese of claim 1 wherein the specific length is two.

3. The method for speech-based information retrieval in Mandarin Chinese of claim 1 wherein the specific length is three.

4. The method for speech-based information retrieval in Mandarin Chinese of claim 1 wherein the indexing terms also can be overlapping character segments with a specific length, and the specific length can be assigned arbitrarily and is at least one.

5. The method for speech-based information retrieval in Mandarin Chinese of claim 1 wherein the indexing terms also can be overlapping word segments with a specific length, and the specific length can be assigned arbitrarily and is at least one.

6. A method for speech-based information retrieval in Mandarin Chinese, comprising:

entering voice or text queries describing information to be requested; determining the indexing terms; and

using the indexing terms to retrieve the information records requested in a format of voice or text type,

wherein the indexing terms are syllable pairs separated by at least one syllable.

7. The method for speech-based information retrieval in Mandarin Chinese of claim 4 wherein the indexing terms also can be character pairs separated by at least one character.

8. The method for speech-based information retrieval in Mandarin Chinese of claim 4 wherein the indexing terms also can be word pairs separated by at least one word.

9. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5; 6, 7 or 8 wherein the selected indexing terms can be of more than one type.

10. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5, 6, 7 or 8 wherein the indexing terms can be one or more types selected from a group comprising overlapping syllable segments, syllable pairs, overlapping character segments, overlapping word segments, character pairs, and word pairs.

11. The method for speech-based information retrieval in Mandarin Chinese of claim 1, 4, 5, 6, 7 or 8 wherein after determining indexing terms, the method for speech-based information retrieval in Mandarin Chinese further comprises:

identifying voice utterances for each syllable, character, or word in the voice queries to generate at least one syllable, character, or word candidate to create corresponding syllable-, character-, or word-lattices; and

identifying voice utterances for each syllable, character, or word in the voice information records to generate at least one syllable, character, or word candidate to create corresponding syllable-, character-, or word-lattices;

each syllable, character, or word candidate of the syllable-, character-, or word-lattices mentioned above comprises a voice recognition score generated by the voice recognition process.

12. The method for speech-based information retrieval in Mandarin Chinese of claim 11 wherein each of the indexing terms further comprises a score, and the score is obtained from averaging the voice recognition scores of all syllable, character, or word candidates involved in the indexing terms.

13. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5, 6, 7 or 8 wherein speech-based information retrieval in Mandarin Chinese comprises using voice query to retrieve text information records, using text query to retrieve voice information records, and using voice query to retrieve voice information records.

14. The method for speech-based information retrieval in Mandarin Chinese of claim 13 wherein the scores of the indexing terms are frequency counts for the indexing terms in the text-type queries or information records if the queries or information records are text-type.

15. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5, 6, 7 or 8 further comprising designing a set of feature vectors for each query and each information record, wherein each feature vector comprises a plurality of components, and each

component is used to represent the scores obtained from the voice recognition process (if in voice-type) or frequency counts (if in text-type) for each indexing term for the queries and information records.

16. The method for speech-based information retrieval in Mandarin Chinese of claim 15 wherein a relationship between the queries and each information record is determined by weighted sum of respective matching results for each pair of corresponding feature vectors representing the query and the information record.

17. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5, 6, 7 or 8 further comprising generating a set of data-driven indexing terms, the generation process for the set of indexing terms can start from a set consisting of all single syllables, characters, or words only in a bottom-up procedure, any two syllable, character, or word segments which appear adjacently iteratively concatenated into a new larger syllable, character, or word segment, if they satisfy some statistical criteria.

18. The method for speech-based information retrieval in Mandarin Chinese of claim 17 wherein the length of the other larger syllable, character, or word segment is two.

19. The method for speech-based information retrieval in Mandarin Chinese of claim 17 wherein the length of the other larger syllable, character, or word segment is three.

20. The method for speech-based information retrieval in Mandarin Chinese of claim 17 wherein the statistical criteria can be mutual information between the two smaller syllable, character, or word segments which appear adjacently and can be concatenated into another larger syllable, character, or word segment.

21. The method for speech-based information retrieval in Mandarin Chinese of claim 17 wherein the statistical criteria can be some

language model parameters between the two smaller syllable, character, or word segments which appear adjacently and can be concatenated into another larger syllable, character, or word segment.

22. The method for speech-based information retrieval in Mandarin Chinese of claim 17 wherein in the step of generating the data-driven indexing terms, when determining whether to combine two adjacent smaller syllable, character, or word segments into a larger syllable, character, or word segment to be new indexing terms, different thresholds are given to the indexing terms of syllable, character, or word segments with different lengths; when the statistical criteria is larger than the threshold, the two smaller syllable, character, or word segments are combined into a new indexing term.

23. The method for speech-based information retrieval in Mandarin Chinese of claim 22 wherein the step of generating the data-driven indexing terms can be performed repeatedly until no statistical criteria of any adjacent syllable, character, or word segments is larger than the threshold.

24. The method for speech-based information retrieval in Mandarin Chinese of claim 11 wherein if the voice recognition score of each syllable, character, or word candidate is smaller than a predetermined value, the syllable, character, or word candidate will be deleted.

25. The method for speech-based information retrieval in Mandarin Chinese of claim 12 wherein if the frequency count of the indexing term in a database is smaller than a predetermined value, the syllable, character, or word candidate will be deleted.

26. The method for speech-based information retrieval in Mandarin Chinese of claim 25 wherein the predetermined value can be set while determining the indexing terms, and different values can be set for different indexing terms.

27. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5, 6, 7 or 8 further comprising creating a list of stop terms according to an inverse document frequency of each indexing term.

28. The method for speech-based information retrieval in Mandarin Chinese of claim 27 further comprising deleting the most frequently occurring indexing terms in the list of stop terms from the feature vectors.

29. The method for speech-based information retrieval in Mandarin Chinese of claims 1, 4, 5, 6, 7 or 8 further comprising creating a term association matrix for the set of indexing terms, the matrix comprising a plurality of matrix elements, each matrix element representing the statistical characteristics for any two indexing terms co-occurring in the same information records.

30. The method for speech-based information retrieval in Mandarin Chinese of claim 29 wherein the matrix elements can be a value between 0 and 1.

31. The method for speech-based information retrieval in Mandarin Chinese of claim 30 wherein the elements equal to 0 may represent two indexing terms never co-occurring in the same information records, or without synonymity association.

32. The method for speech-based information retrieval in Mandarin Chinese of claim 30 wherein the elements equal to 1 may represent two indexing terms always co-occurring in the same information records or with high synonymity association.

33. The method for speech-based information retrieval in Mandarin Chinese of claim 32 further comprising adding the several indexing terms with the highest synonymity association with the indexing terms

in the existing feature vectors of the queries to form new feature vectors of the queries.

34. The method for speech-based information retrieval in Mandarin Chinese of claim 1, 4, 5, 6, 7, 8, 12 or 14 further comprising a second retrieval after the first step of using the indexing terms to retrieve voice- or text-type information records to be requested.

35. The method for speech-based information retrieval in Mandarin Chinese of claim 34 wherein the second retrieval is performed by adding indexing terms or removing indexing terms or modifying their scores to generate new feature vectors of the queries.

36. The method for speech-based information retrieval in Mandarin Chinese of claim 35 wherein the indexing terms to be added, removed or scores modified can be determined by identifying the indexing terms in the feature vectors for the relevant and irrelevant information records obtained in the previous retrieval.

37. The method for speech-based information retrieval in Mandarin Chinese of claim 36 wherein the indexing terms are added or their scores are increased if the indexing terms often appear in the relevant information records obtained in the previous retrieval.

38. The method for speech-based information retrieval in Mandarin Chinese of claim 36 wherein the indexing terms are removed or their scores are decreased if the indexing terms often appear in the irrelevant information records obtained in the previous retrieval.

39. The method for speech-based information retrieval in Mandarin Chinese of claim 11 further comprising a second retrieval after the first step of using the indexing terms to retrieve voice- or text-type information records to be requested.

40. The method for speech-based information retrieval in Mandarin Chinese of claim 39 wherein the second retrieval is performed by adding indexing terms or removing indexing terms or modifying their scores to generate new feature vectors of the queries.

41. The method for speech-based information retrieval in Mandarin Chinese of claim 40 wherein the indexing terms to be added, removed or scores modified can be determined by identifying the indexing terms in the feature vectors for the relevant and irrelevant information records obtained in the previous retrieval.

42. The method for speech-based information retrieval in Mandarin Chinese of claim 41 wherein the indexing terms are added or their scores are increased if the indexing terms often appear in the relevant information records obtained in the previous retrieval.

43. The method for speech-based information retrieval in Mandarin Chinese of claim 41 wherein the indexing terms are removed or their scores are decreased if the indexing terms often appear in the irrelevant information records obtained in the previous retrieval.

44. The method for speech-based information retrieval in Mandarin Chinese of claim 15 further comprising a second retrieval after the first step of using the indexing terms to retrieve voice- or text-type information records to be requested.

45. The method for speech-based information retrieval in Mandarin Chinese of claim 44 wherein the second retrieval is performed by adding indexing terms or removing indexing terms or modifying their scores to generate new feature vectors of the queries.

46. The method for speech-based information retrieval in Mandarin Chinese of claim 45 wherein the indexing terms to be added, removed or scores modified can be determined by identifying the indexing terms

in the feature vectors for the relevant and irrelevant information records obtained in the previous retrieval.

47. The method for speech-based information retrieval in Mandarin Chinese of claim 46 wherein the indexing terms are added or their scores are increased if the indexing terms often appear in the relevant information records obtained in the previous retrieval.

48. The method for speech-based information retrieval in Mandarin Chinese of claim 46 wherein the indexing terms are removed or their scores are decreased if the indexing terms often appear in the irrelevant information records obtained in the previous retrieval.